Storage Devices

Computers are used to process large volumes of data and to execute very complex programs. The computers need to have some kind of storage device to hold these programs and data. Such a device should be directly accessible to the CPU and its speed must be compatible with the speed of the CPU. Also a computer must be able to store frequently needed data on some permanent storage device. Based on the characteristics of the storage devices we classify these devices as Main Memory or secondary storage device. In this chapter we will study the need and use of basic storage devices used with the computers.

4.1 Main Memory

Digital computers are stored-program computers that means a program to be executed is first loaded in the memory and then instructions are executed one by one. The data and results of calculations are also stored in the memory. In this sense main memory is the working area of the computer. It is very fast but limited in capacity. A computer cannot work without having some kind of main memory. Most general purpose computers have enough memory to store a few million characters. In this section we will learn about types of main memory, their use and working principles.

The main memory of a computer consists of thousands or even millions of cells,each capable of storing a bit i.e, 0 or 1. These cells are logically organized into group of 8 bits (**B**inary digi**ts**) called a byte as shown in the Figure 4.1.

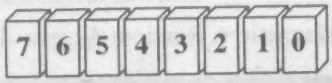


Figure 4.1: Memory cells organized as a byte

Each byte in the memory has a unique number assigned to it. This number is called the address of that byte. This scheme of arranging cells into a byte and bytes into memory chip is shown in Figure 4.2. The number shows the byte number assigned to the byte and is also called its address.

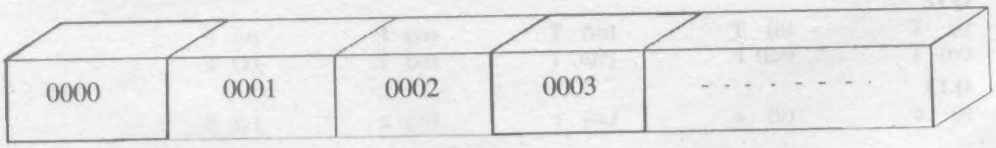


Figure 4.2: Memory addresses

We can view the main memory of a computer as a collection of bytes arranged in an order or sequence. CPU or any other component of the computer can access any byte from the main memory by specifying its address. Different bytes of the main memory can be accessed directly in a random order in equal amount of time. Because of this characteristic of the main memory it is called **direct access storage device**. Accessing any byte of the memory is very fast as compared to other storage devices like the magnetic and optical disks. Most computers have the following two types of main memory.

4.1.1 RAM (Random Access Memory)

RAM is the primary storage device and the data and instructions are stored temporarily in it. It takes the same amount of time to access any location in RAM. CPU can perform two types of operations on RAM, these are:

- Read
- Write

During **Read operation** the contents of memory location are copied to a CPU register whereas during **Write operation** the contents of a CPU register are copied to the memory location. The CPU can not perform any other operations on memory locations. RAM is usually built by using two different technologies i.e. DRAM (Dynamic RAM) and SRAM (Static RAM).

DRAM is the most commonly used technology to build RAM chips and consumes a lot of power as data stored in a DRAM needs to be refreshed periodically.

SRAM is faster than the DRAM but it is more expensive. Unlike DRAM, the contents of SRAM do not need to be refreshed periodically.

In most computers SRAM technology is used to build very fast memory inside a CPU chip. This memory is known as the **cache memory**. Cache memory usually is very small in size as compared to the total memory in the computer but it increases the performance of a computer. This memory arrangement is shown in the Figure 4.3. Following are the main characteristics of RAM :

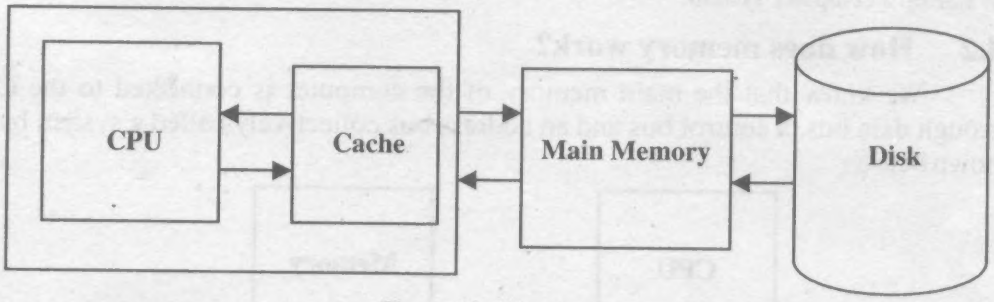


Figure 4.3: Memory management

- The contents of the memory are lost when the electricity supply is cut-off so the main memory is **volatile**.
- Since CPU can read data from and write data to the RAM therefore RAM is read/write memory.
- RAM is random access in the sense that any part of RAM can be accessed directly.

4.1.2 ROM (Read Only Memory)

As is obvious from the name the contents of ROM can be read but new data can not be written into it so it is a **R**ead **O**nly **M**emory. The manufacturer of the ROM writes the data and programs permanently into it and this data and programs can not be changed afterwards. ROM is used to save frequently used instructions and data. The data stored in ROM will not change for a very long time. Following are the commonly used forms of ROM.

4.1.3 PROM (Programmable Read Only Memory)

This form of ROM is initially blank and the user can write his own data/programs on it by using special devices. Once the program/data is written on PROM it cannot be changed or altered. It is obvious that this kind of ROM will be used for storing data for a very long period of time. The data written on this kind of ROM can not be changed once it is written.

4.1.4 EPROM (Erasable Programmable Read Only Memory)

Like PROM it is initially blank and programs and data can be written on it by the manufacturer or by the user with special devices. Unlike PROM a user by using special purpose devices and ultraviolet rays can erase the data written on it. So data/program written on it can be changed and new data can also be added on this form of ROM. As the data written on this kind of ROM can be changed so data that may need to be updated can be written on it but frequently changing data is not written on EPROM.

4.1.5 EEPROM (Electrically Erasable Programmable Read Only Memory)

This kind of ROM can be re-written by using electrical devices and so data stored on EEPROM can be easily modified. EEPROM can be very useful for taking backup of data and for keeping records that are updated periodically.

It is important to note that all the forms of ROM described above are non-volatile so the data stored in these chips is not lost when electricity is cut-off. Mostly ROM chips are used to store frequently used programs like operating system routines (small programs) and data, which is not changed for long periods of time. It is also used to store programs needed to startup a computer system.

4.2 How does memory work?

We know that the main memory of the computer is connected to the CPU through data bus, a control bus and an address bus collectively called a system bus is shown below:

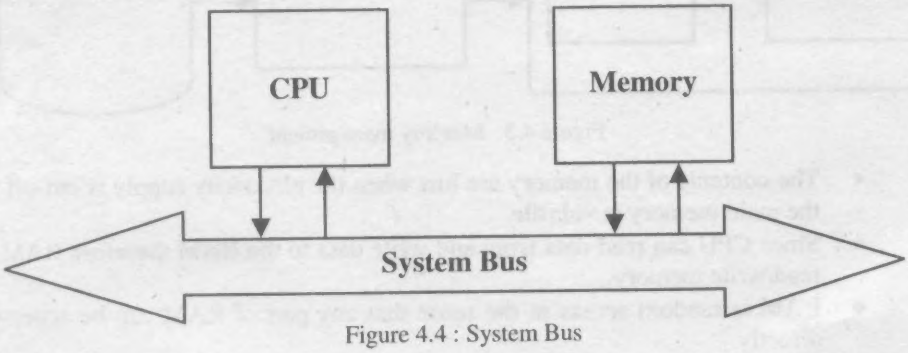


Figure 4.4 : System Bus

When the CPU wants to read data from the memory it first places the read request on the control bus, and also places the address of the byte or word needed on the address bus. The memory unit reads the command and the address and puts the required data on the data bus. The CPU then reads this data from the data bus. Similarly for writing data, CPU first places the Write request on the control bus and also places the address of the word where it wants to write on the Address bus. When memory unit gets ready to do the operation the CPU puts the data on the Data bus and memory unit reads this data and places it in the required word.

As the main memory consists of electronic circuits so a word or byte address is accessible without using any mechanical components. Because of this property the access speed of memory is very fast. Also the data stored in a computer's main memory can be processed in any order. Because of this property the main memory is often referred to as Random Access Memory (RAM). As the RAM is constructed from integrated circuits so it needs to have continuous electrical power supply in order to maintain. When power is switched off all the data stored into it is lost so we say that RAM is volatile.

4.3 Memory Measuring Units

In digital computers the data is represented as a collection of bits. A **bit** is the smallest unit of data that can be used by a computer. We also know that this data is grouped into bytes and a byte is the number of bits needed to store a character. A **byte** is comprised of eight bits. The size of a computer's main memory is often measured as the number of bytes in it. Following is a list of different memory measuring units:

1 Nibble	**= 4 bits**	
1 Byte	**= 8 bits**	
1 KB (Kilo Byte)	**= 1024 bytes**	**$= 2^{10}$ bytes**
1 MB (Mega Byte)	**= 1024 KB**	**$= 2^{20}$ bytes**
1 GB (Giga Byte)	**= 1024 MB**	**$= 2^{30}$ bytes**
1 Terabyte	**= 1024 GB**	**$= 2^{40}$ bytes**

4.4 Data Organization Within a byte or Word

We view the bits within a byte or word as being arranged in a row from left to write. We call one end of this row the high-order end and the other the low-order end. The bit at left end is often called the high-order bit or the Most Significant Bit **(MSB)** ;similarly,the bit at the right end is referred to as the low-order bit or the Least Significant Bit1 **(LSB)**.This is shown in Figure 4.5

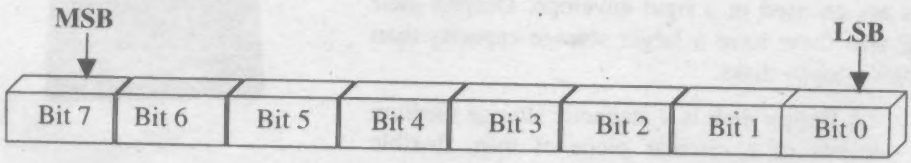


Figure 4.5: A byte with the most significant and least significant bits

4.5 Secondary Memory

Primary memory is directly accessible to the processor and is used to store data and programs that are in current use. The control unit does not have direct access to data that is stored anywhere outside the processor or main memory. However, this storage is limited in size and volatile. We need some storage device that is not temporary in nature and that does not have the same restrictions of size as that of main memory. Such a device is called secondary storage device. Secondary storage devices are categorized according to

- means by which the data is stored, optically or magnetically
- the technique used for storage of the data, sequential storage or direct access storage
- the capacity of the medium, how much can be stored on it
- portability of the medium, can it be moved around easily
- access times to the data stored.

Secondary storage is required to permanently store information that is not needed in memory all of the time and which may be too large to fit into the memory of the computer. Two main categories, based on the ways of accessing data from a secondary storage device, are: **Sequential-Access** and **Direct-Access** or **Serial-Access** and **Random-Access** respectively.

Different computer application programs need these two types of storage devices. For example, a program for calculating the payroll of a company has to access all the data on all of a company's employees, it accesses this data one record at a time, one after the other, and this is called sequential access. Direct access storage device can be used in a departmental store where details of all of the items for sale are needed in a random order. Following table shows a comparison of main memory and secondary memory.

Primary memory	Secondary memory
Expensive	Cheap
small capacity	Large capacity
Connects directly to the processor	Not connected directly to the processor
Fast Access	Slow Access

4.5.1 Floppy disk

Floppy disks are mostly used for transferring data between computer systems and for casual backup of data. They have low capacity, and are very, very slow as compared to other storage devices. The most common size is 3.5 inches diameter. These disks are encased in a rigid envelope. Despite their small size these have a larger storage capacity than the older floppy disks.

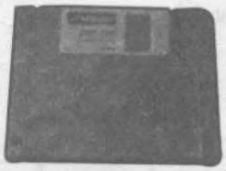


Figure 4.6: Floppy disk

A **floppy disk** is a magnetic storage medium that consists of a circular piece of thin, flexible (hence the name floppy) magnetic media encased in a square or rectangular plastic wallet. Floppy disks are read and written by a floppy disk drive or FDD.

Unlike most **hard disks,** the floppy disks are portable. Floppy disks are slower to access than hard disks and have less storage capacity, but they are much less expensive. Floppies come in three basic sizes i.e. 8-inch, 5¼-inch, 3½-inch, but the last one is the most commonly used.

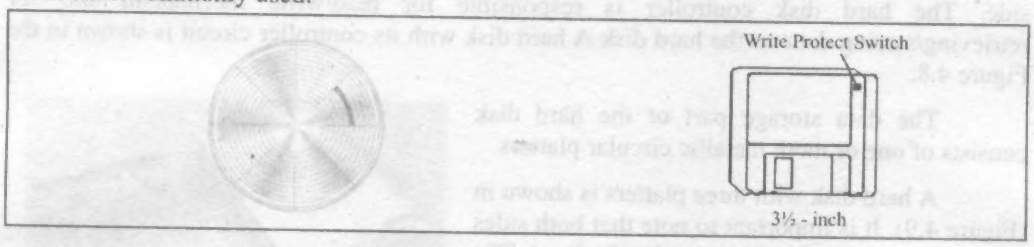


Figure 4.7: Floppy disk – internal view

The following is the series of actions taking place when data is written on the disk

- The computer program passes an instruction to the computer hardware to write a data file on a floppy disk.
- The computer hardware and the floppy-disk-drive controller start the motor in the diskette drive to spin the floppy disk.
- A second motor, called a stepper motor, rotates a worm-gear shaft in minute increments that match the spacing between tracks.
- The read/write heads stop at the track. The read head checks the prewritten address on the formatted diskette to be sure it is using the correct side of the diskette and is at the proper track.
- Then the data is written to the required address.
- The diskette stops spinning. The floppy disk drive waits for the next command.
- On a typical floppy disk drive, the small indicator light stays on during all of the above operations.

4.5.2 Hard Disk

Most digital computers use atleast one hard-disk drive. Some large scale computers normally contain hundreds of hard disks. Hard disks are used to permanently store digital data so you can say that hard disks give computers the ability to remember things when the power goes out. In this section we shall learn the function of a hard disk and also analyze the working of a hard disk.

CAPACITY AND PERFORMANCE

Nowadays a typical desktop computer has a hard disk with a capacity of more than 80 gigabytes. Data is stored onto the disk in the form of files. A file is simply a named collection of bytes. The bytes might be the ASCII (American Standard Code for Information Interchange) codes for the characters of a text file, or they could be the instructions of a software application for the computer, or they could be the stored information, or they could be the pixel colors for an image. There are two ways to measure the performance of a hard disk.

Data rate - The data rate is the number of bytes per second that the drive can deliver to the CPU. Rates between 5 and 40 megabytes per second are common.

Seek time - The time used to move the head to the appropriate track after reading the address is called the **seek time**

A typical hard disk consists of a sealed metallic box with controller circuit on one side. The hard disk controller is responsible for read/write mechanism and for retrieving/storing data on the hard disk A hard disk with its controller circuit is shown in the Figure 4.8.

The data storage part of the hard disk consists of one or more metallic circular platters

A hard disk with three platters is shown in (Figure 4.9). It is important to note that both sides of the platter have their own read/write head. The hard disk controller uses these heads to store and retrieve data from the disk. By arranging data on multiple platters the performance of the hard disk increases.

Figure 4.8: Hard disk – internal view

DATA ORGANIZATION

Data is stored on the surface of a platter in **sectors** and **tracks**. Tracks are concentric circles. The tracks are further divided into sectors. As shown in Figure 4.10 the yellow circle is a track and the blue part represent one sector. Typically a track is divided into 8 sectors. A sector usually contains a fixed number of bytes of data i.e. 512 bytes. When data is to be retrieved from the hard disk the operating system of the computer usually reads the whole track into the memory even if only one byte is needed. This usually increases the performance of the computer system.

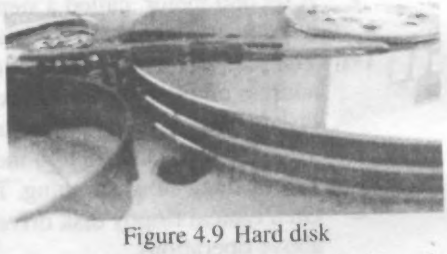


Figure 4.9 Hard disk

As we learned earlier that a hard disk can have more than one platter and each platter have two surfaces. The Tracks on a surface are numbered from 0,1,2 …n . All the tracks on the disk with same track number make up a cylinder.

It is important to note that the position of tracks and sectors are not fixed but these positions are marked by a process called format. Format is of two different types:

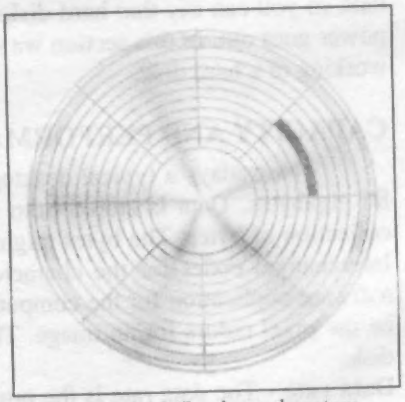


Figure 4.10 Tracks and sectors

LOW LEVEL FORMATTING

During the process of low-level formatting, a drive marks the tracks and sectors on the disk. Usually this is done by the manufacturer of the disk. In this process the starting and ending points of each sector are written onto the platter. This process prepares the drive to hold data.

HIGH LEVEL FORMATTING

During high-level formatting, the information about file-storage is written onto the disk called file-allocation table. This process also prepares the drive to hold data.

HOW DATA IS STORED ON/RETRIEVED FROM THE HARD DISK

As shown above the data is organized into tracks and sectors. Each track has a unique number. First track always has the number 0 0 0 called track zero. Similarly sectors on a track are numbered. When some software or operating system of the computer wants to read some data on some part of the disk it specifies the address of the location and provides the data. By using the provided address, the disk controller moves the read/write heads to the required track. It also uses the motor in the disk to rotate the disk platters. Because of this mechanical component this process is very slow as compared to the speed of the processor. When the head reaches the required track the read/write head has to wait for some time so that the required sector comes under it due to the rotation of the platters. This delay is called the **rotational delay**. When the appropriate sector comes under the read/write head it reads the data from the disk and send this data to the processor. The time consumed in this process is called the **transfer delay**. These three delays are used to calculate the access time of data.

Access Time = Seek Time + Rotational Delay + Transfer Delay

Obviously the seek time and rotational delay involve mechanical parts and are very large. Because of the delays the hard disk is very slow as compared to the CPU.

4.5.3 Compact Disks

One of the most prominent optical storage systems is the Compact Disk (CD), which is compatible with those in the music industry except that computer CD players spin the CD faster than the original CD's used in the music industry to obtain higher data transfer rates.

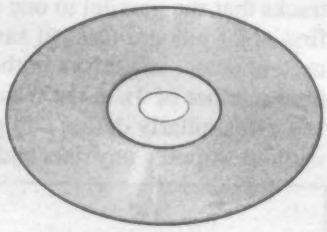


Figure 4.11

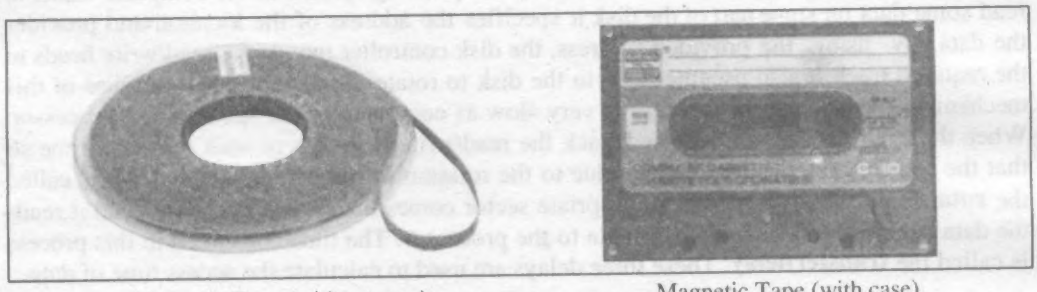
These disks are approximately 5 inches in diameter and consist of reflective material covered with a clear protective coating. Information is recorded on them by creating variations in their reflective surfaces. The information can then be retrieved by detecting these variations with a laser beam. Information on a CD is stored on one continuous track that spirals around the CD like a groove in an old-fashioned record. This is different from the magnetic disks where data is stored in concentric tracks.

The CD is commonly used to store data. CD is usually called CD-ROM (compact disk read-only memory). it can store more than 700 MB of data and are very useful for storing audio and video data. Following is a list of areas where the CD-ROM is used successfully for different purposes.

- Incorporate video on your CD-ROM to make an effective sales tool.
- Distributing different software products e.g., most operating systems are distributed on CD-ROM.
- Distributing audio and video data.
- Keeping the backup of large volumes of data and document archives
- Storing large volumes of data for uses in online application.

4.5.4 Tape Storage

This is an old form of mass storage device that uses magnetic tape. In a magnetic tape information is recorded on the magnetic coating of a plastic tape. To access the data, this tape is mounted in a device called a tape drive that can read, write, and rewind the tape . Tape drives have different sizes ranging from very small cartridge units that use tape similar in appearance to that in stereo systems to large reel-to-reel units. The capacities of these devices vary a lot and some tapes can hold several gigabytes of data.



Magnetic Tape (without case) Magnetic Tape (with case)

Figure 4.12:

HOW DATA IS ORGANIZED ON A MAGNETIC TAPE

Modern streaming tape systems divide a tape into segments, each of which is magnetically marked by a gap when we format the disk. Each of these segments contains several tracks that run parallel to one another lengthwise on the tape this is shown in Figure 4.13. The first eight bits are used to save data and the last bit is used to store parity bit. This bit is used to detect any errors in the data stored on the tape. If this bit is set to 1 or zero so that the total number of 1s in the frame is even. This method of detecting error is called even parity. We can similarly define odd parity. The inter block gaps are needed so that the tape can stop without skipping any data and can be accelerated before reading data.

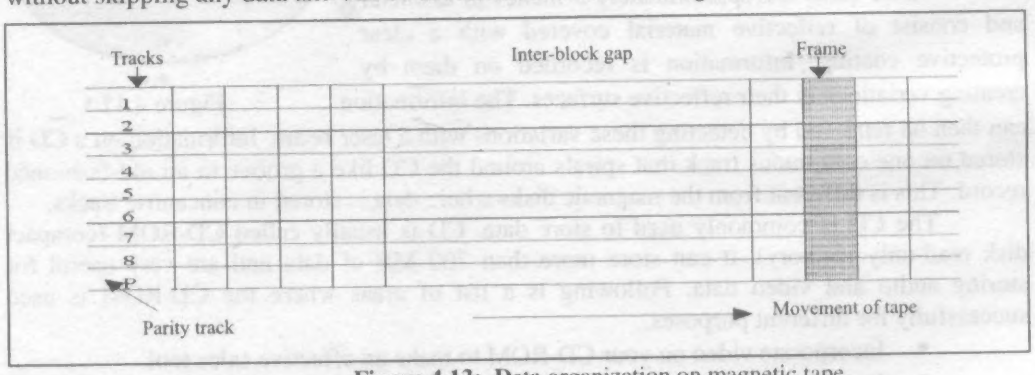


Figure 4.13: Data organization on magnetic tape

Data stored on a magnetic tape can be accessed only sequentially. This is the main disadvantage of streaming tape systems because moving between different positions on a tape can be very time-consuming. Thus tape systems have much longer data access times than disk systems in which different sectors can be accessed by short movements of the read / write

head. The tape systems are not popular for on-line data storage. On the other hand these tape devices are very cheep as compared to the magnetic disks. A large volumes of data can be stored on the tapes for backup purposes so these are used mainly in off-line, backup storage applications.

Exercise

1. Describe in detail the purpose and working of the main memory.
2. Describe in detail the purpose and working of the following Secondary storage devices.
   a. Floppy disk
   b. Hard Disk
3. Describe in detail the purpose and working of the following backing storage devices.
   a. Compact Disks
   b. Magnetic Tape
4. Explain, using a labeled diagram, the concepts of track and sector when describing magnetic disk storage.
5. Explain the purpose of the following and draw a diagram showing their relationship.
   a. Cache memory
   b. Hard disk
   c. Magnetic Tape
6. Explain why secondary memory is needed in a computer system?
7. Explain the purpose of following:
   i. High level formatting
   ii. Low level formatting
   iii. RAM and ROM
8. A 9th class student has a home computer system. What storage devices, the student will use on the home computer system. Explain why these devices are needed?
9. Fill in the blanks :
   (i) _____ is a direct access storage device.
   (ii) _____ is a serial access storage device.
   (iii) Access time = _____ time + _____ time .
   (iv) RAM stands for _____.
   (v) 1 MB is equal to _____bytes.
   (vi) The contents of _____ must be refreshed periodically.
   (vii) The time required to move the head of the hard disk to appropriate track is called _____.
   (viii) The larger the size of the RAM, the _____ the efficiency of the computer.
   (ix) EPROM stands for _____ .
   (x) MSB stands for _____.

head. The tape systems are not popular for on-line data storage. On the other hand these tape devices are very cheep as compared to the magnetic disks. A large volumes of data can be stored on the tapes for backup purposes so these are used mainly in off-line, backup storage applications.

Exercise

1. Describe in detail the purpose and working of the main memory.
2. Describe in detail the purpose and working of the following Secondary storage devices.
   a. Floppy disk
   b. Hard Disk
3. Describe in detail the purpose and working of the following backing storage devices.
   a. Compact Disks
   b. Magnetic Tape
4. Explain, using a labeled diagram, the concepts of track and sector when describing magnetic disk storage.
5. Explain the purpose of the following and draw a diagram showing their relationship.
   a. Cache memory
   b. Hard disk
   c. Magnetic Tape
6. Explain why secondary memory is needed in a computer system?
7. Explain the purpose of following:
   i. High level formatting
   ii. Low level formatting
   iii. RAM and ROM
8. A 9^{th} class student has a home computer system. What storage devices, the student will use on the home computer system. Explain why these devices are needed?
9. Fill in the blanks:
   (i) _____ is a direct access storage device.
   (ii) _____ is a serial access storage device.
   (iii) Access time = _____ time + _____ time .
   (iv) RAM stands for _____.
   (v) 1 MB is equal to _____bytes.
   (vi) The contents of _____ must be refreshed periodically.
   (vii) The time required to move the head of the hard disk to appropriate track is called _____.
   (viii) The larger the size of the RAM, the _____ the efficiency of the computer.
   (ix) EPROM stands for _____.
   (x) MSB stands for _____.

10. Match the following :

Hard disk	Serial access
RAM	Secondary storage device
Tape storage	Optical storage
CD	Primary storage

11. Choose the correct answer

a. Tape storage is

(i) Slower than the hard disk. **(ii)** Faster than hard disk.

(iii) Direct access device. **(iv)** All above.

b. 1 KB is equal to

(i) 1000 bytes **(ii)** 2^{10} bytes **(iii)** 2^{20} bytes **(iv)** 2^{30} bytes

c. Cache memory is

(i) Faster than the main memory. **(ii)** Slower than the main memory.

(iii) Smaller than the main memory. **(iv)** Only(i) and (iii). **(v)** None of the above.

d. Impact printers

(i) Touch the surface of the paper during printing process.

(ii) Don't touch the surface of the paper during printing process.

(iii) Faster than non impact printers. **(iv)** All of the above.

e. Static RAM

(i) contents are required to be refreshed periodically.

(ii) contents are not required to be refreshed periodically.

(iii) is faster than DRAM. **(iv)** only(i)and (ii). **(v)** only(ii) and (iii)

12. Mark the following as True/False

(i) Tape storage is a direct access storage device **(ii)** ROM is volatile

(iii) SRAM is faster than DRAM. **(iv)** 1 MB = 2^{20} bytes

(v) every program must be loaded into RAM before execution

(vi) Zero sector of floppy disk consists of seven tracks

(vii) System Bus is the pathway among different components of computer

(viii) Secondary storage is cheaper than primary storage

(ix) CD is a magnetic storage device

(x) In floating point number format,the exponent is represented in signed magnitude form.

Answers

Q.9

(i) Hard Disk (ii) Magnetic Tape (iii) Seek, Latency (iv) Random Access Memory
(v) 2^{20} (vi) DRAM (vii) Seek Time (viii) Higher
(ix) Erasable Programmable Read Only Memory (x) Most Significant Bit

Q.11

a. i b. ii c. iv d. i e. v

Q.12

(i) F (ii) F (iii) T (iv) T (v) T
(vi) F (vii) T (viii) T (ix) F (x) F